OPEN ACCESS

KING'S College LONDON

King's Research Portal

DOI:
10.1038/ng.3307

Document Version
Peer reviewed version

Link to publication record in King's Research Portal

Citation for published version (APA):
Danjou, F., Zoledziewska, M., Sidore, C., Steri, M., Busonero, F., Maschio, A., Mulas, A., Perseu, L., Barella, S., Porcu, E., Pistis, G., Pitzalis, M., Pala, M., Menzel, S., Metrustry, S., Spector, T. D., Leoni, L., Angius, A., Uda, M., ... Cucca, F. (2015). Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels. *Nature Genetics*, *47*(11), 1264–1271. https://doi.org/10.1038/ng.3307

Citing this paper
Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights
Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

•Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
•You may not further distribute the material or use it for any profit-making activity or commercial gain
•You may freely distribute the URL identifying the publication in the Research Portal

Take down policy
If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 05. May. 2023

nature genetics

Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels

Fabrice Danjou[1,13], Magdalena Zoledziewska[1,13], Carlo Sidore[1–3], Maristella Steri[1], Fabio Busonero[1,2,4], Andrea Maschio[1,2,4], Antonella Mulas[1,3], Lucia Perseu[1], Susanna Barella[5], Eleonora Porcu[1–3], Giorgio Pistis[1–3], Maristella Pitzalis[1], Mauro Pala[1], Stephan Menzel[6], Sarah Metrustry[7], Timothy D Spector[7], Lidia Leoni[8], Andrea Angius[1,8], Manuela Uda[1], Paolo Moi[5,9], Swee Lay Thein[6,10], Renzo Galanello[5,9,12], Gonçalo R Abecasis[2,14], David Schlessinger[11,14], Serena Sanna[1,14] & Francesco Cucca[1,3,14]

We report genome-wide association study results for the levels of A1, A2 and fetal hemoglobins, analyzed for the first time concurrently. Integrating high-density array genotyping and whole-genome sequencing in a large general population cohort from Sardinia, we detected 23 associations at 10 loci. Five signals are due to variants at previously undetected loci: *MPHOSPH9*, *PLTP-PCIF1*, *ZFPM1 (FOG1)*, *NFIX* and *CCND3*. Among the signals at known loci, ten are new lead variants and four are new independent signals. Half of all variants also showed pleiotropic associations with different hemoglobins, which further corroborated some of the detected associations and identified features of coordinated hemoglobin species production.

© 2015 Nature America, Inc. All rights reserved.

The provision of oxygen to tissues depends on hemoglobin, requiring the coordinated expression of several globin chains that form functional tetramers. An index of the importance of hemoglobin function is the evolutionary duplication and divergence of regulation of globin gene copies to adapt to stages of development and buffer the effects of mutational loss. In particular, at birth, a switch occurs from fetal hemoglobin (HbF) to hemoglobin A2 (HbA2) and hemoglobin A1 (HbA1), such that during adult life the hemoglobin forms comprise ~1% HbF, ~3% HbA2 and ~96% HbA1. The different hemoglobins all contain α-globin chains, encoded by two eponymous genes (*HBA1* and *HBA2*) on chromosome 16. These chains aggregate with non–α-globin chains encoded by *HBG1* and *HBG2* (for γ-globin; HbF), *HBD* (for δ-globin; HbA2) and *HBB* (for β-globin; HbA1) genes in the β-globin gene cluster on chromosome 11 (**Fig. 1**). The molecular switch between fetal and adult hemoglobin occurs via the binding of transcription factors to regulatory DNA sequences controlling the expression of globin genes. In particular, the various genes in the β-globin cluster are sequentially activated during ontogeny, such that time-specific expression patterns follow the genomic order[1].

Inherited disorders of hemoglobin, such as β-thalassemia, which is caused by mutations at the *HBB* (hemoglobin β) locus, represent the most common monogenic disorders worldwide[2]. The prevalence of these disorders is highest in areas where malaria was or remains endemic[3]. The severity of inherited hemoglobin disorders is also variable, from severe, lifelong transfusion-dependent anemia to mild anemia that does not require transfusion, depending on the molecular defect and genotype status as well as ameliorating variants in modifier genes. Therefore, studying the genetic regulation of hemoglobin levels might identify new factors involved and mechanisms to optimize strategies for therapy.

The large heritable contribution to phenotypic variance in HbA2 and HbF in the general population (0.728 and 0.633, respectively; Online Methods and a previous report[4]) indicates that genetic analyses could lead to new insights. In genome-wide association studies (GWAS), two genomic regions, the β-globin gene cluster and the *HBS1L-MYB* locus, have been associated at a genome-wide significance level with variations in the amount of HbA2 (ref. 5), and only these loci and *BCL11A* have been associated with HbF levels[6,7]. Variants at all four loci are powerful modifiers of the severity of β-thalassemia and sickle-cell disease[7–10]. Notably, none of the variants associated with HbA2 or HbF levels have been found to be associated with total hemoglobin levels, even in the largest meta-analysis of over

[1]Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche (CNR), Monserrato, Cagliari, Italy. [2]Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, USA. [3]Dipartimento di Scienze Biomediche, Università degli Studi di Sassari, Sassari, Italy. [4]DNA Sequencing Core, University of Michigan, Ann Arbor, Michigan, USA. [5]Ospedale Regionale per le Microcitemie, ASL8, Cagliari, Italy. [6]Department of Molecular Hematology, King's College London, London, UK. [7]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. [8]Center for Advanced Studies, Research and Development in Sardinia (CRS4), Parco Scientifico e Tecnologico della Sardegna, Pula, Italy. [9]Department of Public Health and Clinical and Molecular Medicine, University of Cagliari, Cagliari, Italy. [10]Department of Hematological Medicine, King's College Hospital National Health Service (NHS) Foundation Trust, London, UK. [11]Laboratory of Genetics, National Institute on Aging, US National Institutes of Health, Baltimore, Maryland, USA. [12]Deceased. [13]These authors contributed equally to this work. [14]These authors jointly supervised this work. Correspondence should be addressed to F.D. (fabrice.danjou@irgb.cnr.it) or F.C. (fcucca@uniss.it).

Received 9 December 2014; accepted 23 April 2015; published online 14 September 2015; doi:10.1038/ng.3307

135,000 individuals[11]. This indicates that, in analyses of total hemoglobin levels, the association signals for the subtypes are diluted and possibly obscured by opposite directions of effect. Currently, most of the heritability for variance in HbF and HbA2 levels also remains to be explained, and variation in HbA1 levels has not been specifically assessed by GWAS.

A promising source of genetic data to extend analyses is the founder Sardinian population, in which associations have previously been detected in a large cohort through the analysis of data from genotyping arrays bearing common and ubiquitous variants[7]. Here we extend these analyses to rare and Sardinian-specific variants inferred from whole-genome sequencing of 2,120 Sardinians (**Supplementary Fig. 1** and **Supplementary Note**). Furthermore, analyzing variants modulating HbA1, HbA2 and HbF levels concurrently in a single cohort provides a route to assess associations that overlap for different hemoglobin forms without the need to account for differences in study size, ancestry or measurements.

RESULTS

To test for genetic associations with the levels of HbA1, HbA2 and HbF, we interrogated ~10.9 million SNPs, genotyped or imputed in 6,602 general population volunteers of the SardiNIA longitudinal study[4] (Online Methods and **Supplementary Table 1**).

Initial analyses showed a predominant role for the *HBB* c.118C>T mutation introducing a premature stop codon (p.Gln40*), better known as the $\beta^0$39 mutation, a variant common in Sardinia (rs11549407: allele frequency = 4.8%). This mutation results in complete absence of β-globin chain synthesis (β^0) and consequent β-thalassemia in homozygous individuals and results in a decrease in HbA1 levels and an increase in HbA2 and HbF levels in heterozygous individuals ($P < 1.0 \times 10^{-200}$). Because its effect had been established previously[7,12], we considered this mutation and other rarer β^0 mutations in β-thalassemia known in Sardinia as covariates (Online Methods and **Supplementary Table 2**). The assessed individuals in the cohort included 664 healthy heterozygous carriers of β^0 mutations but no individuals with β-thalassemia.

The genome-wide scan identified 23 unique variants at 10 loci at the classical significance threshold of $P = 5 \times 10^{-8}$. Of note, 21 loci were significant even when considering a more stringent threshold of $P = 1.4 \times 10^{-8}$, calculated on the basis of an empirical estimate of the number of independent tests in the Sardinian genome (see the companion paper[13]).

Five variants were at previously undetected loci, four variants were new, independent signals at known loci and ten variants refined previously described associations to new lead polymorphisms that may have functional effects (**Table 1**). We observed 6, 14 and 8 independent genome-wide significant signals for HbA1, HbA2 and HbF levels, respectively (**Supplementary Fig. 2**). Hence, some of the associated variants significantly affected more than one hemoglobin, resulting in 28 variant-trait associations (**Fig. 2**, **Table 1** and **Supplementary Table 3**). Variants resulting from imputation and not supported by data for linked genotyped markers were experimentally validated (**Supplementary Table 4**).

New associations at newly associated loci

New associations were detected for all three hemoglobin forms. For HbA1, we observed a signal led by chr12:123681790 (in an intron of *MPHOSPH9*), encompassing several SNPs in complete linkage disequilibrium (LD) in a region with several genes (**Supplementary Fig. 3**). Which gene is truly associated and how it affects hemoglobin production remain unclear, although, among the top associated SNPs, a variant in an intron of *ARL6IP4* (chr12:123465483) falls in a highly conserved region rich in putative transcription factor binding sites and had the highest score for *in silico* prediction of deleterious impact on function (combined annotation-dependent depletion (CADD) score)[14], as detailed in **Supplementary Table 5**. Although this association was just below the more stringent empirical threshold of significance, it was further strengthened by independent association with another hemoglobin form (HbA2, $P = 5.9 \times 10^{-5}$) (**Table 1**).

For HbA2, we identified three new signals. One, rs141006889, is a missense variant located in *ZFPM1*, a gene also known as *FOG1* that encodes a cofactor of the hematopoietic transcription factors GATA1 and GATA2 (ref. 15; **Supplementary Fig. 4**). The complexes formed by FOG1 and GATA proteins are essential for normal erythroid differentiation[15], as demonstrated by pathogenetic mutations that abrogate the FOG1-GATA interaction to cause familial dyserythropoietic anemia and thrombocytopenia[16]. Another signal was defined by a pair of statistically indistinguishable variants, rs113267280 and rs112233623 ($P = 1.11 \times 10^{-29}$ and 1.29×10^{-29}, respectively),

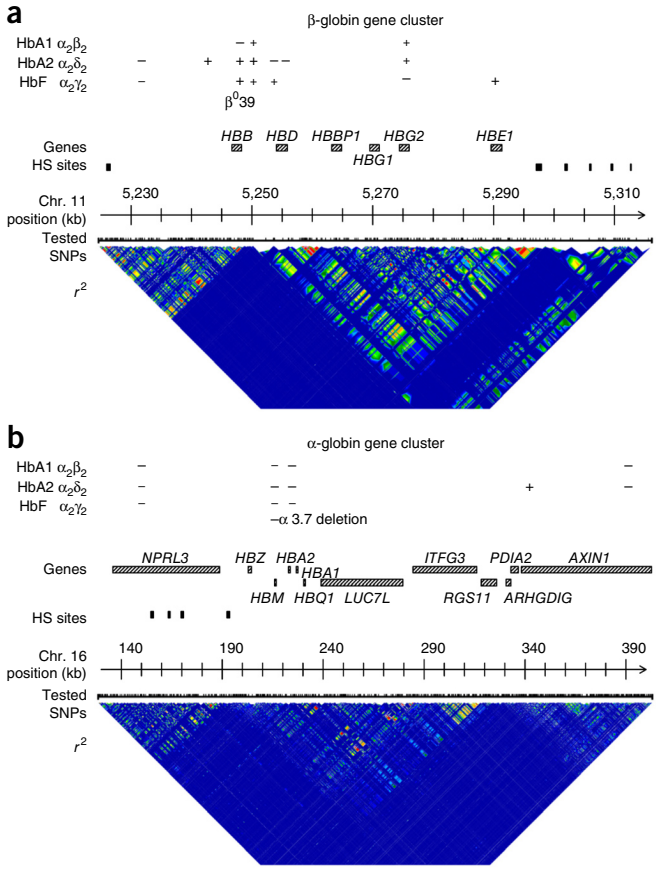

Figure 1 Association at the globin gene clusters. (**a**,**b**) Schematics of the association results in the genomic context of the β-globin (**a**) and α-globin (**b**) gene clusters. For each hemoglobin form, the positions of the associated markers are indicated, with plus and minus signs indicating an increase or decrease in the levels of the corresponding hemoglobin form with the effect allele (as in **Table 1**). The symbol is larger if the marker is associated at the genome-wide level or smaller if the association results from the analysis of pleiotropic effects. The $\beta^0$39 mutation and −α 3.7 type I deletion as well as relevant genes and the locus control region hypersensitivity (HS) sites are indicated. Finally, at the bottom of each panel is presented the LD (r^2) profile for the region in Sardinia, with colors representing a range of LD from high (red) to intermediate (green) to low (blue).

© 2015 Nature America, Inc. All rights reserved.

located in the *CCND3* gene, whose product, cyclin D3, is thought to be critical for erythropoiesis[17]. Knockdown of *Ccnd3* correlates with a reduction in the number of cell divisions during terminal erythropoiesis, thereby resulting in the production of fewer red blood cells that are larger in size[18]. These variants are also in partial LD with rs9349205 (r^2 = 0.40), a SNP previously associated with mean red blood cell volume and number (**Supplementary Table 6**), which is located 160 bp away from rs112233623 in the same

Table 1 Most significant independent association results from single-variant tests for A1, A2 and fetal hemoglobin

Trait (units) and locus number	Candidate gene	Chromosome: position	rsID from dbSNP 142	Alleles (EA/OA)	RSQR	EAF	Effect (SE)	*P* value	Shared effects		
									HbA1	HbA2	HbF
HbA1 (g/dl)											
Locus 1[a]	α-globin gene cluster (p, o, b); *MPG* (p)	16:149539[a,d,f]	rs570013781	A/G	0.98	0.136	−0.1995 (0.023)	5.86×10^{-18}	–	–	–
	α-globin gene cluster (p, o, b); *AXIN1* (p)	**16:391593 (cond.)[a,c,e]**	**–**	**T/C**	**0.94**	**0.012**	**−0.4028 (0.058)**	**3.28×10^{-12}**	**–**	**–**	
Locus 2	*FAM3A* (p); *G6PD* (p, c, o, b); *IKBKG* (p)	X:153762634[d,f]	rs5030868	A/G	Genotyped	0.085	−0.1256 (0.019)	2.78×10^{-11}	–		
Locus 3[b]	***MPHOSPH9* (p)**	**12:123681790[b]**	**–**	**A/C**	**0.96**	**0.010**	**−0.3606 (0.064)**	**1.68×10^{-8}**	**–**	**–**	
HbA2											
Locus 1 (%)[d]	β-globin gene cluster (p, o, b); *HBD* (c)	11:5255582[d,f]	rs35152987	A/C	Genotyped	0.004	−2.182 (0.109)	4.35×10^{-86}		–	
	β-globin gene cluster (p, o, b); *HBD* (c)	11:5251849 (cond.)[d,f]	rs7944544	T/G	0.98	0.005	−1.26 (0.097)	3.90×10^{-38}		–	+
	β-globin gene cluster (p, o, b); *HBB* (c); *HBG1-HBG2* (e); *OR51V1* (p)	11:5231565 (cond.)[d,f]	rs12793110	T/C	1.00	0.181	−0.2408 (0.019)	5.75×10^{-36}		–	–
	β-globin gene cluster (p, o, b); *OR51V1* (p)	11:5242698 (cond.)[d,f]	rs11036338	C/G	0.99	0.381	0.1282 (0.017)	2.03×10^{-14}		+	
	β-globin gene cluster (p, o, b); *HBG1-HBG2* (e)	11:5250168 (cond.)[d,f]	rs7936823	G/A	0.96	0.466	0.1117 (0.015)	5.00×10^{-13}	+	+	+
Locus 2 (g/dl)[a,c,e]	α-globin gene cluster (p, o, b); *HBM* (c); *LUC7L* (p)	**16:216593[a,c]**	**rs141494605**	**C/T**	**0.97**	**0.149**	**−0.3080 (0.025)**	**3.94×10^{-35}**	**–**	**–**	**–**
	α-globin gene cluster (p, o, b); *AXIN1* (p)	**16:391593 (cond.)[a,c,e]**	**–**	**T/C**	**0.94**	**0.012**	**−0.5112 (0.063)**	**6.48×10^{-16}**	**–**	**–**	
	α-globin gene cluster (p, o, b); *ARHGDIG* (p); *AXIN1* (p); *ITFG3* (p); *PDIA2* (p); *RGS11* (p)	**16:342218 (cond.)[a,c,e]**	**rs148706947**	**T/C**	**0.93**	**0.021**	**0.2892 (0.051)**	**1.04×10^{-8}**		**+**	
Locus 3 (%)[b]	***CCND3* (p, b)**	**6:41952511[b]**	**rs113267280**	**G/T**	**0.99**	**0.101**	**0.2923 (0.026)**	**1.11×10^{-29}**		**+**	**+**
Locus 4 (%)	*MYB* (b)	6:135418916	rs7776054	G/A	Genotyped	0.210	0.1762 (0.020)	3.71×10^{-19}		+	+
Locus 5 (%)[b]	***CTSA* (p); *PCIF1* (p, c); *PLTP* (p, e); *MMP9* (e); *TNNC2* (e)**	**20:44547672[b]**	**rs59329875**	**C/T**	**1.00**	**0.134**	**−0.1399 (0.024)**	**3.64×10^{-9}**		**–**	
Locus 6 (%)[b]	***FOG1* (p, b, c); *C16orf85* (p)**	**16:88601281[b]**	**rs141006889**	**G/A**	**Genotyped**	**0.007**	**−0.5074 (0.087)**	**5.33×10^{-9}**		**–**	
HbF (g/dl)											
Locus 1	*BCL11A* (p, o, b)	2:60720951	rs4671393	A/G	1.00	0.136	0.578 (0.023)	2.60×10^{-130}			+
	BCL11A (p, o, b)	2:60710571 (cond.)[d,f]	rs13019832	A/G	1.00	0.484	−0.2024 (0.017)	9.12×10^{-33}			–
Locus 2	*MYB* (b)	6:135419018	rs9399137	C/T	Genotyped	0.205	0.4202 (0.020)	1.09×10^{-93}		+	+
	HBS1L (p, c, e); *ALDH8A1* (e)	**6:135356216 (cond.)[d]**	**rs11754265**	**C/G**	**1.00**	**0.367**	**−0.1421 (0.021)**	**5.04×10^{-12}**			**–**
Locus 3[d]	β-globin gene cluster (p, o, b); *HBG1-HBG2* (e)	11:5290370[d]	rs67385638	G/C	1.00	0.236	0.2038 (0.019)	1.09×10^{-25}			+
	β-globin gene cluster (p, o, b); *HBG1-HBG2* (e)	11:5277236 (cond.)[d,f]	rs2855122	C/T	1.00	0.395	−0.1458 (0.022)	2.57×10^{-11}	+	+	–
Locus 4[b,e]	***NFIX* (p)**	**19:13121899[b,e]**	**rs183437571**	**T/C**	**0.97**	**0.010**	**0.4607 (0.081)**	**1.61×10^{-8}**			**+**

The table shows the most significant association results (all results are corrected for β^0 mutations observed in the *HBB* gene, and results for the α-globin gene cluster are adjusted for –α 3.7 deletion type I; Online Methods). New signals are shown in bold. At each locus, we indicate the chromosome and genomic position (hg19 build), the rsID when available, the effect allele tested for association (EA) and the other allele at the SNP (OA), the imputation accuracy (RSQR), the SNP effect allele frequency (EAF) and the regression coefficients (with standard error, SE). We indicate when a SNP is also linked to the other hemoglobin forms ($P < 0.01$) and specify the direction of the effect for the allele indicated in the EA column (+, increases hemoglobin levels; –, decreases hemoglobin levels). The candidate genes likely to be modulated by the lead SNP are also reported along with their inclusion criteria, as described in the Online Methods (p, position; c, coding; e, eQTL; o, Online Mendelian Inheritance in Man (OMIM); b, biological). When the α-globin gene cluster is mentioned, we are referring to the *NPRL3*, *HBZ*, *HBQ1*, *HBA1*, *HBA2* and *HBM* genes; when the β-globin gene cluster is mentioned, we are referring to the *HBB*, *HBD*, *HBBP1*, *HBG1*, *HBG2* and *HBE1* genes. Association coefficients for males and females separately are reported in **Supplementary Table 11**. Cond., results obtained by conditional analysis including as covariates the markers from the row(s) above for the considered locus.

[a]Association results locally corrected for –α 3.7 deletion type I (g.34164_37967del3804, NG_000006.1) (**Supplementary Note**). [b]First time associated with the trait and in a new locus. [c]First time associated with the trait in a previously reported locus. [d]Signal refinement at a previously reported signal. [e]Result not found using the 1000 Genomes Project reference panel. [f]Variant refining previously reported signals.

© 2015 Nature America, Inc. All rights reserved.

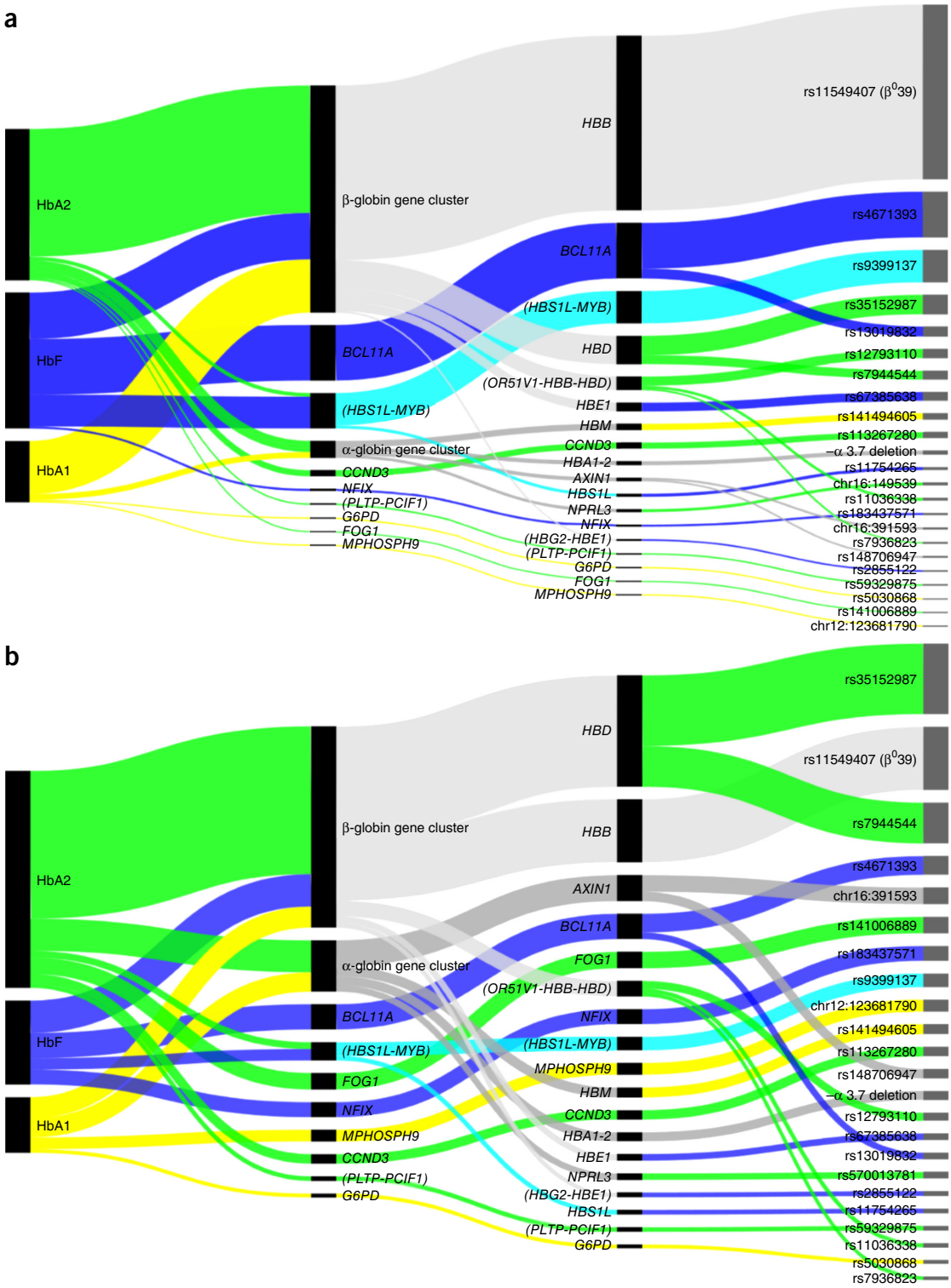

Figure 2 Diagrams of genome-wide associated loci. (**a**,**b**) Representations of genome-wide significant findings for hemoglobin levels in relation to their contribution to the phenotypic variation (variance explained) (**a**) and their individual impact (effect size) (**b**). At each level, the length of the black bar represents the magnitude of the variance explained (**a**) or the effect size (**b**) for each trait, locus, gene and variant. The bars are connected by colored bands to their subcomponents (loci for each trait, genes for each locus, variants for each gene). Three colors (yellow, green and blue) represent the three hemoglobin forms (HbA1, HbA2 and HbF, respectively); for loci or genes affecting more than one hemoglobin, gray indicates effects shared by HbA1 and HbA2, cyan indicates effects shared by HbA2 and HbF, and light gray indicates effects shared by all three hemoglobin forms. Each panel is drawn to show loci in order of their importance, from the largest to smallest amount of explained phenotypic variance (**a**) or effect size (**b**). The variance explained by each locus was calculated by fitting a regression model including all variants at that locus, and the effect size for each locus is the sum of the effect sizes of all the variants in that locus (**Supplementary Table 3** reports the effect sizes for such joint models). For variants associated with more than one trait, the maximum value was used. Markers are reported using chromosome:position notation when an rsID was not available; when an intergenic region was involved, we show nearby genes in parentheses instead of indicating a single gene for the locus.

© 2015 Nature America, Inc. All rights reserved.

erythroid-specific enhancer functionally associated with *CCND3* (refs. 18–20). rs112233623 was also the associated variant with the highest CADD score (**Supplementary Table 5**).

An additional variant related to HbA2, rs59329875, was observed for the first time in this study. It is situated between *PLTP*, which has been associated with several plasma lipoprotein and triglyceride levels[21–24], and *PCIF1*, which is thought to negatively regulate gene expression by RNA polymerase II (ref. 25).

For HbF, we identified one new variant associated with its levels: rs183437571, located on chromosome 19 in an intron of *NFIX*, which encodes a CCAAT-binding transcription factor. The association for this variant was just below the empirical significance threshold of $P = 1.4 \times 10^{-8}$ but is supported by considerable biological evidence implicating the gene and the surrounding region in hemoglobin regulation. Specifically, rs183437571 falls in a CpG region that is differentially methylated in fetal and adult red blood cell progenitors[26]. In mice, *Nfix* was recently identified as one of the regulatory factors with relatively restricted expression in hematopoietic stem cells[27] and was required for the survival of hematopoietic stem and progenitor cells during stress hematopoiesis[28]. Intriguingly, *NFIX* is situated in a region of ~300 kb that encompasses a number of genes involved in erythropoiesis (*DNASE2* and *KLF1*)[29–33] or otherwise associated with red blood cell traits, including mean corpuscular hemoglobin levels (*SYCE2*, *FARSA* and *CALR*)[11] (**Supplementary Fig. 5** and **Supplementary Table 6**). *KLF1* is a particularly interesting candidate gene[33,34], but the mutations observed in previous studies[35] were not found and the gene itself is situated in an LD block distinct from the one corresponding to our association signal. However, long-distance regulatory interactions remain a possibility.

Of the five new signals, the discovery of chr12:123681790 for HbA1, rs141006889 for HbA2 and rs183437571 for HbF was strongly influenced by the assessment of variants from Sardinian whole-genome sequencing. Specifically, chr12:123681790 was missing from 1000 Genomes Project phase 3 data[36], and, when using this public reference panel, the signal was misplaced at another variant ~1 Mb away; rs141006889 was included in the design of one genotyping array (ExomeChip) after it was identified through our sequencing effort but is currently not detected in sequenced 1000 Genomes Project samples; and rs183437571 was poorly imputed with 1000 Genomes Project phase 3 data, with a resulting signal that was not genome-wide significant (**Table 1** and **Supplementary Table 7**).

Overall, the amount of variance explained by markers associated at the genome-wide level (**Table 1**) accounted for a fraction of the estimated genetic component for each trait (ranging from 46% for HbA1 to 68% for HbA2; Online Methods), supporting inheritance models that include variants that have small effect sizes and/or are rare. For instance, 21 additional genes with signals showing suggestive levels of significance ($P < 1 \times 10^{-4}$, minor allele frequency (MAF) >0.5%) were related to the genome-wide significant loci listed here, either in the scientific literature (PubMed records before 2006) or by expression levels (Human Expression Atlas[37]) or Gene Ontology[38] categories, as determined using GRAIL software[39] (**Supplementary Table 8** and **Supplementary Note**). Four of the suggestive signals most strongly linked to genome-wide association findings were located in *NFE2*, which encodes erythroid nuclear factor 2 (ref. 40); *ADGB*, which encodes a recently discovered globin of unknown physiological function[41]; and *SPTB* and *ANK1*, both of which encode proteins affecting the stability of erythrocyte membranes[42].

To test for replication of the associations at new loci detected in Sardinians, we used the largest independent sample reported thus far, which measured HbA2 and HbF levels as well as F cell percentage (the proportion of erythrocytes containing HbF) (Online Methods) in 4,131 individuals from the TwinsUK cohort enrolled from the UK general population[43]. For two loci, both associated with HbA2 levels, we successfully replicated the associations seen in Sardinia. Specifically, we observed a *P* value of 6.98×10^{-6} for rs59329875 in the *PLTP-PCIF1* intergenic region (MAF = 0.18) and a *P* value of 1.73×10^{-4} for rs113267280 in *CCND3* (MAF = 0.01). The rarity of other variants precluded replication. The *MPHOSPH9* and *FOG1* variants associated with HbA1 and HbA2 levels, respectively, are missing in publicly available imputation panels, and rs183437571 in *NFIX*, which associated with HbF levels in Sardinia, was imputed as monomorphic in the TwinsUK cohort (**Table 2** and Online Methods).

Fine mapping at known loci

The integration of variants from whole-genome sequencing in the scan was also instrumental in refining signals at previously known loci, either identifying a better lead variant or indicating new independent signals. Specifically, we refined the associations within the α- and β-globin gene clusters for all three hemoglobins; the association of the *HBS1L-MYB* intergenic region with HbA2 and HbF; and the association of the *BCL11A* gene with HbF.

Table 2 Replication of new loci

Trait (units) and locus number	SNP	Candidate gene	INFO score	Alleles (EA/OA)	EAF	Effect (SE)	*P* value	Notes
HbA1 (g/dl)								
Locus 3	chr12:123681790	*MPHOSP9*	_	_	_	_	_	Not imputable because absent in the 1000 Genomes Project; at the moment, Sardinian specific
HbA2 (%)								
Locus 3	rs113267280	*CCND3*	0.843	G/T	0.011	0.442 (0.118)	1.73×10^{-4}	.
Locus 5	rs59329875	*PLPT-PCIF1*	0.994	C/T	0.185	0.132 (0.029)	6.98×10^{-6}	
Locus 6	rs141006889	*FOG1*	_	_	_	_	_	Not imputable because absent in the 1000 Genomes Project; detected in the NHLBI GO Exome Sequencing Project (ESP)
HbF (%)								
Locus 4	rs183437571	*NFIX*	0.294	T/C	0.000	_	_	Imputed as monomorphic in the TwinsUK cohort

The table describes association in the TwinsUK cohort ($n = 4,131$ individuals). For each SNP, we indicate the associated hemoglobin tested, the imputation accuracy according to the IMPUTE INFO metric, the effect allele tested for association (EA) and the other allele at the SNP (OA), the SNP effect allele frequency (EAF) and the regression coefficients (with standard error, SE). The last column explains why some SNPs were not tested. The information in the first column is presented as in **Table 1**.

© 2015 Nature America, Inc. All rights reserved.

The associations within the β-globin gene cluster were intricate. As reported above, the strongest modifier in this region is the *HBB* $\beta^0$39 variant, which acts on all three hemoglobin types (**Fig. 1**, Online Methods and **Supplementary Table 2**). Multiple additional independent signals were observed in conditional analyses for HbA2 and HbF, but they were distinct for each hemoglobin type, highlighting different regulatory patterns within the β-globin gene cluster. Specifically, for HbA2, we confirmed two known independent associations at missense mutations in the *HBD* gene (rs35152987 and rs35406175; the latter was perfectly tagged by our lead signal; **Supplementary Table 5**). In addition, we identified three new independent signals (rs12793110, rs11036338 and rs7936823) within an LD block encompassing the *HBB* gene, confirming a controlling role for this region in HbA2 production[5] (**Fig. 1** and **Supplementary Fig. 4**). For HbF, two new independent signals were detected in a separate LD block of the β-globin gene cluster (**Fig. 1** and **Supplementary Fig. 5**). The first, situated in an intron of the *HBE1* gene (rs67385638), remained associated even when taking into account 43 other variants in the β-globin gene cluster associated with hemoglobin variation (**Supplementary Note**). The second was located in a cyclic AMP response element upstream of *HBG2* (rs2855122) already implicated in drug-mediated HbF induction by butyrate[44]: the different features of this marker make it a strong candidate in modulating fetal to adult hemoglobin switching (**Supplementary Note**).

At the α-globin gene cluster, two variants were associated with HbA1 and three variants were associated with HbA2, of which one affected both traits (**Fig. 1** and **Table 1**). All results at this locus were corrected for any effect of the most frequent α-globin gene deletion present in Sardinia (g.34164_37967del3804 (NG_000006.1), known as –α 3.7 deletion type I), directly genotyped in a subset of the volunteers and imputed for the rest of the cohort (Online Methods). This deletion was associated at the genome-wide level with both HbA1 and HbA2 levels but only nominally with HbF levels (**Table 1** and **Supplementary Table 2**). The most strongly associated signals (rs570013781 and rs141494605) were situated within the *NPRL3* and *HBM* genes, affecting HbA1 and HbA2 levels, respectively. *NPRL3* contains several hypersensitive sites involved in regulation of the α-globin genes. *HBM* encodes a globin member of the avian α-D family[45]; its expression is highly regulated in human erythroid cells, although the protein has not been detected in human erythroid tissues. These observations suggest a possible regulatory function for which high-level protein expression would not be required[45]. An independent variant associated with HbA1 and HbA2 (chr16:391593) was observed in the *AXIN1* gene, in which a further independent SNP (rs148706947) was found to be associated with HbA2 alone (**Supplementary Figs. 3** and **4**).

We also examined variants in the *HBS1L-MYB* intergenic region known to be associated with HbF and HbA2 levels[5]. We confirmed the role of the known variant (rs66650371, a TAC deletion) on the expression of both forms of hemoglobin[46,47] (**Supplementary Note**). A further new independent signal for HbF was found at rs11754265 in an intron of *HBS1L*, which has been shown to be a much stronger expression quantitative trait locus (eQTL) than rs66650371 for *HBS1L* and the neighboring *ALDH8A1* gene in monocytes[48].

In line with previous studies[6–8,49,50], the second intron of *BCL11A* gave multiple signals of association with HbF levels. These can be explained by the joint action of variants from each of two independent groups of statistically indistinguishable SNPs: one group constitutes rs4671393, rs766432 and rs1427407, with P values between 2.6×10^{-130} and 5.6×10^{-129}, and the other group constitutes rs13019832 and rs7606173, with P values of 6.1×10^{-33} and 9.1×10^{-33}, respectively, in our cohort. The most likely causal candidate in the first group is rs1427407, a variant already associated with HbF levels in other population cohorts and functionally associated with *BCL11A* regulation[51]. In the second group, we can point to rs13019832 as the candidate causal variant, as this SNP had the highest functional CADD score (**Supplementary Table 5**). This variant has also been correlated in adipose tissue with the methylation of a CpG site (cg23678058) in a region that is functionally associated with *BCL11A* expression[52] and shows evidence of an effect on GATA-1 binding in peripheral blood–derived erythroblasts[53,54].

Pleiotropic effects

Among our 23 lead variants, 6 were associated (with at least $P < 0.01$) with a second hemoglobin type, and another 6 (including $\beta^0$39 and –α 3.7 deletion type I) were associated with all 3 hemoglobin forms (**Fig. 1** and **Table 1**). Overall, all but three pleiotropic variants modulated different hemoglobins in the same manner, that is, with the same allele increasing the levels of all associated hemoglobins. The three exceptions included the $\beta^0$39 variant, which decreased HbA1 levels while increasing HbA2 and HbF levels, and two SNPs mapping to the β-globin gene cluster, both affecting HbA2 and HbF levels but in opposite directions (**Fig. 1** and **Table 1**). Moreover, many of the additional suggestive signals were associated with more than one hemoglobin type, increasing the likelihood that they are true signals (Online Methods). In fact, 14 of these variants—all sharing effects on HbA1 and HbA2 levels but none having effects on HbF levels—showed between-trait combined P values that were genome-wide significant (**Supplementary Table 9**) and hint at additional pathways of potential interest in hemoglobin dynamics.

In general, the extended number of genetic variants showing joint association with HbA1 and HbA2 rather than HbF is consistent with high correlations of the levels of the adult hemoglobins HbA1 and HbA2 but only partial correlations of these hemoglobin forms with the levels of HbF (Online Methods).

Given the central role of hemoglobin in providing oxygen to body tissues and the substantial fraction of total body cells accounted for by circulating red blood cells, factors influencing hemoglobin production and red blood cell count unsurprisingly have pleiotropic effects on other non-hematological traits. This is exemplified by the strong impact of the major $\beta^0$39 mutation on cholesterol and low-density lipoprotein (LDL) cholesterol levels (see the companion paper[13]). Here we extended the analysis for this mutation to 69 non-hematological quantitative traits selected from among those assessed in the SardiNIA cohort[4] (**Supplementary Note**). We found that this variant was also significantly associated with increased total white blood cell counts ($P = 3 \times 10^{-7}$), with the major contribution coming from neutrophil counts ($P = 1 \times 10^{-6}$), and platelet counts ($P = 9 \times 10^{-5}$) (**Supplementary Table 10**).

DISCUSSION

We provide evidence for 23 associated variants at 10 loci influencing the levels of one or more of the 3 hemoglobin species measurable in postnatal life. Our results are based on a cohort from the Sardinian founder population that is much larger than the samples in previously described GWAS for HbF and HbA2, and the analysis interrogates a high-resolution genetic map based on population sequencing that expanded the assessed spectrum of allelic variants by tenfold in comparison to previous studies. The finding that two of the five newly reported loci were not detectable without using the SardiNIA reference panel and that the others were mislocalized (**Table 1** and **Supplementary Table 7**) further highlights how large-scale sequencing efforts in this founder population can identify

© 2015 Nature America, Inc. All rights reserved.

functionally relevant variants that may be very rare and hence missed in other populations.

For the same reasons, however, replication of the results for such variants or translation of findings directly into other populations is difficult. For example, the other currently reported sample of comparable size, from the UK, could provide replication only for the two variants present there. Similar limitations will likely be found in other GWAS designed to detect the effects of rare and founder variants. However, additional corroboration of our findings for such variants comes from their independent associations with other hemoglobin species and hematological traits in Sardinians and also from the biological functions of the genes involved. For instance, variant chr12:123681790 in *MPHOSP9*, associated with HbA1 levels, also shows suggestive evidence of association with HbA2 levels. The variant in *FOG1*, which is very rare in Europeans (MAF = 0.4%), is a missense variant in a gene implicated in erythropoiesis, and the variant in *NFIX*, which is absent in other European populations, falls within a cluster of genes involved in erythropoiesis and in a CpG region differentially methylated in fetal and adult red blood cell progenitors[26].

By carrying out GWAS analysis for HbA1, HbA2 and HbF levels assessed, to our knowledge, for the first time in the same individuals, we see a wide range of pleiotropic effects of variants across the three hemoglobin types (**Table 1**). Strikingly, HbA2 is associated with more than half of the loci discovered here (**Fig. 2**), with many of the loci having pleiotropic effects on HbA1 and some having such effects on HbF. Thus, although HbA2 has a minor role in the transport of oxygen to tissues[55], variations in its levels participate in pathways that regulate the amounts of the other hemoglobins active in postnatal life.

The direction of the pleiotropic effects among the different hemoglobin types provides some additional clues to mechanism. Within the α-globin gene cluster, in agreement with the presence of α-globin chains in HbA1, HbA2 and HbF, all variants affecting more than one hemoglobin show the same direction of effect on all. The regulation of globin chains from the β-globin gene cluster, however, is more complicated. It involves variants with the same direction of effect for all hemoglobins (rs7936823) and other variants most likely involved in switching mechanisms that affect fetal and adult hemoglobins in opposite directions (rs2855122). Still other variants change the kinetics of competition among non–α-globin chains; for example, the $\beta^0$39 mutation decreases β-globin levels and thereby increases the availability of α-globin chains for combination with δ- and γ-globins, leading to higher levels of HbA2 and HbF.

Variants influencing only two forms of hemoglobin acted mainly in the same direction and never jointly affected HbA1 and HbF levels. Among these variants, those shared only by HbA2 and HbF can be attributed to specific *cis*-regulatory mechanisms in the β-globin gene cluster (rs12793110 and rs7944544) or to loci with a role in erythroid differentiation (*CCND3* and *MYB*). By contrast, variants shared by HbA2 and HbA1 either acted in *trans* (in *MPHOSPH9*) or were localized to the α-globin gene cluster but with effect sizes probably too small to influence HbF production. Consistent with the latter possibility, –α 3.7 deletion type I, which has strong genome-wide significant effects on HbA1 and HbA2, had much smaller only suggestive effects on HbF (**Supplementary Table 2**).

Our analyses also detected broader pleiotropic effects, most strikingly for the $\beta^0$39 variant. In addition to the effects on LDL cholesterol described in the companion paper[13], we report for the first time, to our knowledge, that the $\beta^0$39 mutation is also significantly associated with increased total counts of white blood cells (and some subsets thereof) as well as platelet counts. These findings suggest that in heterozygous carriers this variant drives a broader increase in bone marrow–derived blood cell counts. Speculatively, some of these effects, such as augmented leukocyte and neutrophil counts, might have provided protection against pathogens in addition to those causing malaria, thus increasing selection for the balanced polymorphism.

The detected variants provide candidate modifiers influencing the clinical status of patients with monogenic hemoglobin disorders. For example, we carried out a preliminary analysis of a small sample of 306 patients with β-thalassemia homozygous for the $\beta^0$39 stop-gain mutation but showing very great heterogeneity in disease presentation and course. In addition to the variants described previously[7–10], some variants detected in this study showed possible effects as modifiers of disease severity (**Supplementary Note**). However, the potential of these variants in helping to predict disease severity remains tentative without studies of larger sample sets. Nevertheless, the variants already add to the candidate targets for therapeutic intervention in the widely prevalent inherited β-thalassemia and other hemoglobinopathies[2].

URLs. SardiNIA project, https://sardinia.irp.nia.nih.gov/; 1000 Genomes Project, http://www.1000genomes.org/; HumanExome BeadChip design, http://genome.sph.umich.edu/wiki/Exome_Chip_Design; Immunochip, http://www.t1dbase.org/poster/the-immuno-chip-custom-genotype-array/; Cardio-Metabochip, http://csg.sph.umich.edu//kang/MetaboChip/; Human OmniExpress BeadChip, http://www.illumina.com/applications/genotyping/human-genotyping-arrays/omni-arrays.html; GenomeStudio software, http://www.illumina.com/applications/microarrays/microarray-software/genomestudio.html; MACH software, http://csg.sph.umich.edu/abecasis/MACH/; Minimac software, http://genome.sph.umich.edu/wiki/Minimac; zCall software, https://github.com/jigold/zCall/; IMPUTE v2 software, http://mathgen.stats.ox.ac.uk/impute/impute_v2.1.0.html; Merlin (including Merlin-regress and Merlin-offline), http://csg.sph.umich.edu/abecasis/merlin/; EPACTS software, http://genome.sph.umich.edu/wiki/EPACTS; Metal software, http://csg.sph.umich.edu/abecasis/metal/; GWAS catalog, http://www.genome.gov/gwastudies/; GRAIL software, https://www.broadinstitute.org/mpg/grail/; Gene Ontology, http://geneontology.org/; Human Expression Atlas at BioGPS, http://biogps.org/; Pritchard eQTL browser, http://eqtl.uchicago.edu/; R project, http://www.r-project.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work is dedicated to Antonio Cao, Renzo Galanello and Maurizio Longinotti, who devoted their scientific lives to understanding, preventing and treating hematological diseases in Sardinia. We are also grateful to M.S. Ristaldi and M.G. Marini for knowledge and insight that they freely shared with us. Finally, we thank all the volunteers who generously participated in this study and made this research possible. The SardiNIA study was funded in part by the US National Institutes of Health (National Institute on Aging, National Heart, Lung, and Blood Institute, and National Human Genome Research Institute). This research was supported by National Human Genome Research Institute grants HG005581, HG005552, HG006513 and HG007022; by National Heart, Lung, and Blood Institute grant HL117626; by the Intramural Research Program of the US National Institutes of Health, National Institute on Aging, contracts N01-AG-1-2109 and HHSN271201100005C; by Sardinian Autonomous Region (L.R. number 7/2009) grant cRP3-154; by grant FaReBio2011 "Farmaci e Reti Biotecnologiche di Qualità"; and by the PB05 InterOmics MIUR Flagship Project. The TwinsUK study was funded by the Wellcome Trust; the European Community's Seventh Framework Programme

© 2015 Nature America, Inc. All rights reserved.

(FP7/2007-2013); and the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. Genotyping in the replication cohorts was performed by the Wellcome Trust Sanger Institute and National Eye Institute via the US National Institutes of Health/Center for Inherited Disease Research (CIDR). S.L.T. was supported by the Medical Research Council, UK (grant G0000111, ID51640), and S. Menzel received funding from the British Society for Haematology (start-up grant).

AUTHOR CONTRIBUTIONS

G.R.A., D.S. and F.C. conceived the study. F.D., D.S., S.S. and F.C. drafted the manuscript. F.D., M.Z., M.U., P.M., S.L.T., G.R.A., D.S., S.S. and F.C. revised the manuscript. F.B., A. Maschio and A.A. performed sequencing experiments. M. Pitzalis, G.R.A. and S.S. selected samples for sequencing. F.D., C.S., M.S., E.P., G.P. and S.S. carried out genetic association analyses in the SardiNIA cohort. C.S. analyzed DNA sequence data. M.Z., F.B. and A. Mulas carried out SNP array genotyping. M.Z. designed the validation strategy, and M.Z., F.B. and A. Mulas verified genotypes by Sanger sequencing and TaqMan genotyping. L.P. performed genotyping of –α 3.7 deletion type I. M. Pala created an automatized pipeline to query the public eQTL repositories. P.M. and R.G. provided genotypes and phenotypic data for patients with β-thalassemia. S.B. and R.G. supervised the characterization of the hemoglobins in the SardiNIA cohort. F.D. analyzed the cohort of patients with β-thalassemia. S. Menzel, T.D.S. and S.L.T. provided replication samples. S. Metrustry analyzed replication samples. L.L. provided IT support for sequencing and genotype data processing and analyses. D.S. and F.C. supervised the study. All authors reviewed and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Sankaran, V.G., Xu, J. & Orkin, S.H. Advances in the understanding of haemoglobin switching. *Br. J. Haematol.* **149**, 181–194 (2010).
2. Modell, B. & Darlison, M. Global epidemiology of haemoglobin disorders and derived service indicators. *Bull. World Health Organ.* **86**, 480–487 (2008).
3. Malaria Genomic Epidemiology Network. Reappraisal of known malaria resistance loci in a large multicenter study. *Nat. Genet.* **46**, 1197–1204 (2014).
4. Pilia, G. *et al.* Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet.* **2**, e132 (2006).
5. Menzel, S., Garner, C., Rooks, H., Spector, T.D. & Thein, S.L. HbA2 levels in normal adults are influenced by two distinct genetic mechanisms. *Br. J. Haematol.* **160**, 101–105 (2013).
6. Bae, H.T. *et al.* Meta-analysis of 2040 sickle cell anemia patients: *BCL11A* and *HBS1L-MYB* are the major modifiers of HbF in African Americans. *Blood* **120**, 1961–1962 (2012).
7. Uda, M. *et al.* Genome-wide association study shows *BCL11A* associated with persistent fetal hemoglobin and amelioration of the phenotype of β-thalassemia. *Proc. Natl. Acad. Sci. USA* **105**, 1620–1625 (2008).
8. Lettre, G. *et al.* DNA polymorphisms at the *BCL11A*, *HBS1L-MYB*, and β-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. *Proc. Natl. Acad. Sci. USA* **105**, 11869–11874 (2008).
9. Danjou, F. *et al.* Genetic modifiers of β-thalassemia and clinical severity as assessed by age at first transfusion. *Haematologica* **97**, 989–993 (2012).
10. Danjou, F. *et al.* A genetic score for the prediction of β-thalassemia severity. *Haematologica* **100**, 452–457 (2015).
11. van der Harst, P. *et al.* Seventy-five genetic loci influencing the human red blood cell. *Nature* **492**, 369–375 (2012).
12. Trecartin, R.F. *et al.* Beta zero thalassemia in Sardinia is caused by a nonsense mutation. *J. Clin. Invest.* **68**, 1012–1017 (1981).
13. Sidore, C. *et al.* Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat. Genet.* doi:10.1038/ng.3368 (14 September 2015).
14. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
15. Freson, K. *et al.* Molecular cloning and characterization of the GATA1 cofactor human FOG1 and assessment of its binding to GATA1 proteins carrying D218 substitutions. *Hum. Genet.* **112**, 42–49 (2003).
16. Nichols, K.E. *et al.* Familial dyserythropoietic anaemia and thrombocytopenia due to an inherited mutation in *GATA1*. *Nat. Genet.* **24**, 266–270 (2000).
17. Kozar, K. *et al.* Mouse development and cell proliferation in the absence of D-cyclins. *Cell* **118**, 477–491 (2004).
18. Sankaran, V.G. *et al.* Cyclin D3 coordinates the cell cycle during differentiation to regulate erythrocyte size and number. *Genes Dev.* **26**, 2075–2087 (2012).
19. Soranzo, N. *et al.* A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat. Genet.* **41**, 1182–1190 (2009).
20. Kamatani, Y. *et al.* Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat. Genet.* **42**, 210–215 (2010).
21. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.* **41**, 56–65 (2009).
22. Jarvik, G.P. *et al.* Genetic and nongenetic sources of variation in phospholipid transfer protein activity. *J. Lipid Res.* **51**, 983–990 (2010).
23. Lettre, G. *et al.* Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project. *PLoS Genet.* **7**, e1001300 (2011).
24. Kettunen, J. *et al.* Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat. Genet.* **44**, 269–276 (2012).
25. Hirose, Y. *et al.* Human phosphorylated CTD–interacting protein, PCIF1, negatively modulates gene expression by RNA polymerase II. *Biochem. Biophys. Res. Commun.* **369**, 449–455 (2008).
26. Lessard, S., Beaudoin, M., Benkirane, K. & Lettre, G. Comparison of DNA methylation profiles in human fetal and adult red blood cell progenitors. *Genome Med.* **7**, 1 (2015).
27. Riddell, J. *et al.* Reprogramming committed murine blood cells to induced hematopoietic stem cells with defined factors. *Cell* **157**, 549–564 (2014).
28. Holmfeldt, P. *et al.* Nfix is a novel regulator of murine hematopoietic stem and progenitor cell survival. *Blood* **122**, 2987–2996 (2013).
29. Kawane, K. *et al.* Requirement of DNase II for definitive erythropoiesis in the mouse fetal liver. *Science* **292**, 1546–1549 (2001).
30. Porcu, S. *et al.* Klf1 affects DNase IIα expression in the central macrophage of a fetal liver erythroblastic island: a non-cell-autonomous role in definitive erythropoiesis. *Mol. Cell. Biol.* **31**, 4144–4154 (2011).
31. Zhou, D., Liu, K., Sun, C.-W., Pawlik, K.M. & Townes, T.M. KLF1 regulates BCL11A expression and γ- to β-globin gene switching. *Nat. Genet.* **42**, 742–744 (2010).
32. Siatecka, M. & Bieker, J.J. The multifunctional role of EKLF/KLF1 during erythropoiesis. *Blood* **118**, 2044–2054 (2011).
33. Satta, S. *et al.* Compound heterozygosity for *KLF1* mutations associated with remarkable increase of fetal hemoglobin and red cell protoporphyrin. *Haematologica* **96**, 767–770 (2011).
34. Borg, J. *et al.* Haploinsufficiency for the erythroid transcription factor KLF1 causes hereditary persistence of fetal hemoglobin. *Nat. Genet.* **42**, 801–805 (2010).
35. Perseu, L. *et al.* *KLF1* gene mutations cause borderline HbA2. *Blood* **118**, 4454–4458 (2011).
36. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
37. Su, A.I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* **101**, 6062–6067 (2004).
38. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
39. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* **5**, e1000534 (2009).
40. Andrews, N.C. The NF-E2 transcription factor. *Int. J. Biochem. Cell Biol.* **30**, 429–432 (1998).
41. Hoogewijs, D. *et al.* Androglobin: a chimeric globin in metazoans that is preferentially expressed in Mammalian testes. *Mol. Biol. Evol.* **29**, 1105–1114 (2012).
42. Iolascon, A., Perrotta, S. & Stewart, G.W. Red blood cell membrane defects. *Rev. Clin. Exp. Hematol.* **7**, 22–56 (2003).
43. Moayyeri, A., Hammond, C.J., Valdes, A.M. & Spector, T.D. Cohort profile: TwinsUK and Healthy Ageing Twin Study. *Int. J. Epidemiol.* **42**, 76–85 (2013).
44. Sangerman, J. *et al.* Mechanism for fetal hemoglobin induction by histone deacetylase inhibitors involves γ-globin activation by CREB1 and ATF-2. *Blood* **108**, 3590–3599 (2006).
45. Goh, S.-H. *et al.* A newly discovered human α-globin gene. *Blood* **106**, 1466–1472 (2005).
46. Farrell, J.J. *et al.* A 3-bp deletion in the *HBS1L-MYB* intergenic region on chromosome 6q23 is associated with HbF expression. *Blood* **117**, 4935–4945 (2011).
47. Stadhouders, R. *et al.* *HBS1L-MYB* intergenic variants modulate fetal hemoglobin via long-range *MYB* enhancers. *J. Clin. Invest.* **124**, 1699–1710 (2014).
48. Zeller, T. *et al.* Genetics and beyond—the transcriptome of human monocytes and disease susceptibility. *PLoS ONE* **5**, e10693 (2010).
49. Bhatnagar, P. *et al.* Genome-wide association study identifies genetic variants influencing F-cell levels in sickle-cell patients. *J. Hum. Genet.* **56**, 316–323 (2011).
50. Bauer, D.E. & Orkin, S.H. Update on fetal hemoglobin gene regulation in hemoglobinopathies. *Curr. Opin. Pediatr.* **23**, 1–8 (2011).
51. Bauer, D.E. *et al.* An erythroid enhancer of *BCL11A* subject to genetic variation determines fetal hemoglobin level. *Science* **342**, 253–257 (2013).
52. Grundberg, E. *et al.* Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am. J. Hum. Genet.* **93**, 876–890 (2013).
53. Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* **42**, D764–D770 (2014).
54. Rosenbloom, K.R. *et al.* ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.* **41**, D56–D63 (2013).
55. Steinberg, M.H. & Adams, J.G. Hemoglobin A2: origin, evolution, and aftermath. *Blood* **78**, 2165–2177 (1991).

© 2015 Nature America, Inc. All rights reserved.

ONLINE METHODS

Sample description. The population studied here included 6,921 individuals, representing >60% of the adult population of 4 villages in the Lanusei valley in Sardinia, Italy. These individuals are part of the SardiNIA project, a longitudinal study including genetic and phenotypic data for 1,257 multigenerational families with more than 37,000 relative pairs. Details of phenotype assessments for these samples have been published previously[4]. All participants gave informed consent to study protocols, which were approved by the Sardinian local research ethic committees: Comitato Etico di Azienda Sanitaria Locale 8, Lanusei (2009/0016600) and Comitato Etico di Azienda Sanitaria Locale 1, Sassari (2171/CE)) and by the NIH Office of Human Subject Research as governed by Italian institutional review board approval.

For whole-genome sequencing, we selected 1,122 individuals from the SardiNIA study and 998 individuals enrolled in case-control studies of multiple sclerosis and type 1 diabetes in Sardinia. Genomes were sequenced to an average coverage of 4.16-fold. Details on the sequencing protocol, data processing and variant calling can be found elsewhere[56] and in the companion paper[13]. The 2,120 sequenced samples consisted of 695 complete and incomplete parent-offspring trios; to avoid over-representation of rare haplotypes during the imputation process, we considered only the parents from each trio—totaling 1,488 samples—when building our reference panel[56] (see the companion paper[13] for further details).

Part of the sequencing data used in this study are available through the database of Genotypes and Phenotypes (dbGaP), under the SardiNIA Medical Sequencing Discovery Project, study accession phs000313.v3.p2.

Genotyping and imputation. The four microarrays used to genotype the entire SardiNIA cohort were the Illumina Infinium HumanExome BeadChip, Immunochip, Cardio-Metabochip and HumanOmniExpress BeadChip. Genotyping was carried out according to the manufacturer's protocols at the SardiNIA Project Laboratory (Lanusei, Italy), at the Technological Center–Porto Conte Ricerche (Alghero, Italy) and at the National Institute on Aging Intramural Research Program Laboratory of Genetics (Baltimore, Maryland, USA). Genotypes were called using GenomeStudio (version 1.9.4) and refined using zCall (version 3)[57]. We applied standard per-sample quality control filters to remove samples with low call rates or for which the reported relationships and/or sex was discordant with the genetic data. The details of the quality control filters are described elsewhere[56]. Altogether, 890,542 autosomal markers and 16,325 X-linked markers were genotyped across the SardiNIA study samples. We selected for phasing and imputation only the 6,602 samples for which genotyping was successfully carried out on all 4 arrays.

The genotypes were phased using MACH software[58], with 30 iterations of the haplotyping Markov chain and 400 states per iteration. We performed imputation using Minimac software[59] and a reference panel including the haplotypes from 1,488 Sardinian whole genomes[56] (see the companion paper[13]). Variants with estimated imputation quality (RSQR) ≤0.3 or <0.8 were discarded if the estimated MAF was ≥1% or 0.5–1%, respectively; variants with MAF <0.5% were kept only if genotyped. The RSQR thresholds for rare and low-frequency variants were more stringent than those proposed for other traits[56] as they led to better genomic control parameters (1.001, 0.993 and 0.985 for HbA1, HbA2 and fetal hemoglobin, respectively). We also performed imputation with the 1000 Genomes Project phase 3 (version 5)[36] haplotype set and used the same thresholds to discard variants. The genomic control parameters for 1000 Genomes Project imputation were 1.050, 0.997 and 0.984 for HbA1, HbA2 and fetal hemoglobin, respectively.

Association analysis. We performed association analyses with the concentrations (g/dl) for all three hemoglobins as well as the percentages for HbA2 and HbF. HbA2 and HbF percentages were directly measured from high-performance liquid chromatography (HPLC), and HbA1, HbA2 and HbF concentrations (g/dl) were derived from total hemoglobin amounts measured by Coulter counter. As expected, the measurements in percentage and grams per deciliter were highly correlated for HbF (Spearman's $\rho = 0.99$) and HbA2 ($\rho = 0.85$). The HbA1 percentage was not considered for genetic association because it was too highly correlated with both the HbA2 and HbF percentages as a consequence of their derivation formulas ($\rho = -0.803$ and -0.757, respectively; $P < 1 \times 10^{-20}$). Considering only non-carriers of β^0 mutations, HbA1 (g/dl) was highly correlated with HbA2 (g/dl) ($\rho = 0.662$; $P < 1 \times 10^{-20}$) and poorly correlated with HbF (g/dl) ($\rho = -0.055$; $P = 3.44 \times 10^{-5}$). Likewise, HbA2 and HbF were weakly positively correlated as percentage measures ($\rho = 0.108$; $P = 4.08 \times 10^{-16}$) and were even less correlated when measured in grams per deciliter ($\rho = 0.066$; $P = 5.81 \times 10^{-5}$), consistent with previous findings[5]. Measurements were available for a subset of 6,305 individuals; descriptive statistics are reported in **Supplementary Table 1**. Association results were considered genome-wide significant when the P value was $<5 \times 10^{-8}$; however, we also note in the text variants that would not meet a threshold of $P = 1.4 \times 10^{-8}$ that we introduced for sequencing-based GWAS carried out in Sardinians for variants with MAF >0.5% (see the companion paper[13]).

Before association analyses, trait measures were normalized using inverse normal transformation; we also removed for all assessed traits outliers with values above 5% for HbF. Analyses were adjusted for age, age^2 and sex as well as for the presence of at least one of the three β^0 mutations ($\beta^0$39 (rs11549407), *HBB* c.20delA (rs63749819) and *HBB* c.315+1G>A (rs33945777)), all directly genotyped or sequenced. Regression coefficients for the $\beta^0$39 variant—the most common in Sardinia, with 10.3% of the population being a carrier—are reported in **Supplementary Table 2**.

Association was performed using the q.emmax test in EPACTS[60], which implements a linear mixed-model procedure to correct for cryptic relatedness and population stratification by incorporating a genomic-based kinship matrix. The associations reported in **Table 1** refer to the best P value obtained with either percentage or original measurement units for HbA2 and HbF. Notably, HbF signals always resulted in lower P values when considering measurements in grams per deciliter, whereas, for HbA2 analysis, this was only the case for rs141494605. All loci passed the genome-wide significance threshold of $P < 5 \times 10^{-8}$ for both percentage and grams per deciliter except for rs59329875, which was genome-wide significant only for the HbA2 measure reported in **Table 1**.

To identify independent signals, we performed regional conditional analysis, using a forward selection procedure that added, at each step, the most associated variant as a covariate in the model. In this sequential analysis, we tested only SNPs lying in a 2-Mb region centered on the lead variant. The same genome-wide significance threshold used for primary signals was also considered for independent signals. For loci where different independent signals were found, we also report the model parameters of the jointly associated variants in **Supplementary Table 3**. Finally, the lead variants and their surrogates (r^2 >0.90) were annotated with CADD scores[14] and are reported in **Supplementary Table 5**.

Heritability and variance explained. We estimated heritability for the three hemoglobins using Merlin-regress[61] on the same sample used for the GWAS. The estimates for normalized hemoglobin levels were 0.520 g/dl for HbA1, 0.728% for HbA2 (0.700 g/dl) and 0.633% for HbF (0.624 g/dl). We then calculated for each hemoglobin form the proportion of phenotypic variance explained by the associated lead variants. We measured this proportion as the difference in the R^2-adjusted parameter estimated for the full and the basic models, where the basic model included only phenotypic covariates (age, age^2 and sex) and the full model also included all the independent SNPs associated with the specific trait. R^2-adjusted values were calculated using a linear mixed-model procedure from the lmekin() function in the kinship R package[62]. The estimates were 0.240 g/dl for HbA1, 0.492% for HbA2 and 0.383% for HbF.

Characterization of β^0 mutations. For the present study, we designed a TaqMan custom assay for the *HBB* c.118C>T nonsense mutation (rs11549407, also known as $\beta^0$39) and genotyped 6,602 samples. Comparison of the TaqMan genotypes and the imputation results (rs11549407, RSQR = 0.92) showed an overall concordance rate of 98.8%. Also, we further sequenced all samples discordant between red blood cell index-based diagnosis (using mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), HbF percentage and HbA2 percentage) and TaqMan genotypes, using Sanger sequencing to determine any additional β-globin mutations different from $\beta^0$39, thus identifying three carriers for the *HBB* c.20delA (rs63749819) mutation and one carrier for the *HBB* c.315+1G>A (rs33945777) mutation.

Characterization of the deletion in the α-globin gene cluster. In Sardinia, three variants are known to be mainly responsible for α-thalassemia: SNPs

© 2015 Nature America, Inc. All rights reserved.

doi:10.1038/ng.3307

rs111033603 and rs41474145 and the deletion g.34164_37967del3804 (NG_000006.1); the latter, known as –α 3.7 deletion type I, is by far the most common[63]. We did not observe the rarer rs111033603 and rs41474145 SNPs in our sequencing effort. To establish genotypes at the deletion site in the full cohort, we used an inference strategy combined with experimental data. Specifically, we first characterized the structural variant by PCR in 260 unrelated sequenced individuals randomly selected from the SardiNIA cohort. We calculated the relative coverage of the deleted region in the whole genome–sequenced samples by considering the ratio of read counts in the potentially deleted region (223,450–226,953 bp, excluding 150-bp boundaries) with read counts in the nearby region not subject to deletion (227,254–230,757 bp). We then identified coverage ratio thresholds that best predicted PCR genotypes for the deletion and used these thresholds to infer genotypes for the 2,120 sequenced individuals. We inserted the genotypes in the Sardinian reference panel and imputed the deletion genotype for the total SardiNIA cohort. To assess the accuracy of imputation, we considered the best-guess genotypes and searched for Mendelian errors in families. The observed error rate was 0.58% over 1,193 parent-offspring pairs, consistent with high imputation precision. The association results reported in the manuscript for this locus are corrected for inferred –α 3.7 deletion type I dosages.

Variant validation. We validated all variants that showed genome-wide significant *P* values in the primary or conditional analysis that were not directly genotyped or had no surrogates (r^2 >0.90) that were directly genotyped. We did not validate variant rs13019832 in *BCL11A* for HbF, which was highly linked with findings from previous reports (rs7606173)[49,51]. Validation was performed using Sanger sequencing or TaqMan genotyping, depending on the variant frequency, for five variants. We selected for each variant all individuals carrying the minor allele (heterozygous and homozygous) plus a random subset of individuals homozygous for the other allele (in all, 3,084 subjects were genotyped), except for rs141494605 and chr16:391593, for which we specifically selected worse imputation dosages (borderline RSQR values). In addition, for rs17525396, we used independent genotypes available for a subset of the cohort[64], derived from Affymetrix 6.0 arrays (**Supplementary Table 4**).

Replication of variant effects. Replication was performed in the TwinsUK cohort[43]. Genotyping was performed using a combination of Illumina arrays (HumanHap300, HumanHap610Q, 1M-Duo and 1.2MDuo 1M), and imputation was performed using the IMPUTE software package (v2) and 1000 Genomes Project haplotypes released on 16 June 2014 in the phase 1 integrated variant set release[36,65]. Details on quality control filters are provided in the **Supplementary Note**. HbA2 concentrations and HbF percentages were obtained by HPLC, and F cells were enumerated by staining for intracellular HbF and subsequent flow cytometry[66]. Measurements were available for 4,131 samples. Association analyses were performed with the Merlin-offline package in Merlin, to account for relatedness[61]. To be consistent with analyses performed in the SardiNIA study, age, age^2 and sex were used as covariates and the trait measures were transformed using quantile normalization.

Selection of candidate genes. At each locus, we generated a list of genes to be considered as plausible candidates, including those that satisfied one of the following criteria: (i) genes that were located within 25 kb of the lead SNP (indicated by p in **Table 1**); (ii) genes with exonic variants (frameshift, stop codon, nonsynonymous or synonymous) along with splice-site and 5′ or 3′ UTR variants in LD ($r^2 \geq 0.8$) with the lead SNP (indicated by c in **Table 1**); (iii) genes whose expression was modulated by the SNP itself or by an eQTL in LD ($r^2 \geq 0.8$) with the top SNP (indicated by e in **Table 1**); (iv) genes with a clear biological function connected with the traits (indicated by b in **Table 1**); or (v) genes harboring variants responsible for Mendelian diseases, as reported in OMIM (indicated by o in **Table 1**). We searched for candidate genes in eQTL data using an automatized pipeline querying 16 public eQTL repositories[48,67–81], including the Pritchard eQTL browser; only top SNP eQTLs or SNPs with a false discovery rate (FDR) <0.05 were considered.

Pleiotropy and gene connection analysis. To characterize genome-wide significant results and to identify suggestively significant ones, we searched for shared effects on the different hemoglobin forms as well as evidence of connections between them. Specifically, for genome-wide significant markers, we have simply reported the effect direction for all traits with $P < 0.01$ when a marker was associated at the genome-wide level for one trait (**Table 1**). To identify candidates with suggestive *P* values between 1.00×10^{-4} and 5.00×10^{-8}, we selected among the following: (i) markers with MAF >0.5% and showing two-trait combined *P* values $<5 \times 10^{-8}$ (*P* values were combined using inverse variance–weighted meta-analysis, as implemented in Metal software[82]) and (ii) markers located in or near genes that demonstrated evidence of connections with genome-wide significant loci, either in PubMed (using a set with data from 2006 or earlier to avoid confounding by subsequent GWAS discoveries) or the Human Expression Atlas[37] and Gene Ontology[38] databases, as determined using GRAIL[39], and considering genes reported with multiple hypothesis–corrected $P < 0.05$. Using these criteria, we identified 21 further genes with biological connections to the genome-wide significant loci, reported in **Supplementary Table 8**, and 14 variants with combined *P* values between 2.08×10^{-8} and 1.18×10^{-11}, reported in **Supplementary Table 9**.

56. Pistis, G. *et al.* Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. *Eur. J. Hum. Genet.* **23**, 975–983 (2015).
57. Goldstein, J.I. *et al.* zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinformatics* **28**, 2543–2545 (2012).
58. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
59. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
60. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
61. Abecasis, G.R., Cherny, S.S., Cookson, W.O. & Cardon, L.R. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30**, 97–101 (2002).
62. R Core Development Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2013).
63. Origa, R. *et al.* Complexity of the α-globin genotypes identified with thalassemia screening in Sardinia. *Blood Cells Mol. Dis.* **52**, 46–49 (2014).
64. Naitza, S. *et al.* A genome-wide association scan on the levels of markers of inflammation in Sardinians reveals associations that underpin its complex regulation. *PLoS Genet.* **8**, e1002480 (2012).
65. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
66. Menzel, S. *et al.* A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat. Genet.* **39**, 1197–1199 (2007).
67. Myers, A.J. *et al.* A survey of genetic human cortical gene expression. *Nat. Genet.* **39**, 1494–1499 (2007).
68. Stranger, B.E. *et al.* Population genomics of human gene expression. *Nat. Genet.* **39**, 1217–1224 (2007).
69. Veyrieras, J.-B. *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* **4**, e1000214 (2008).
70. Dimas, A.S. *et al.* Common regulatory variation impacts gene expression in a cell type–dependent manner. *Science* **325**, 1246–1250 (2009).
71. Pickrell, J.K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
72. Fehrmann, R.S.N. *Trans*-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet.* **7**, e1002197 (2011).
73. Innocenti, F. *et al.* Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genet.* **7**, e1002078 (2011).
74. Montgomery, S.B., Lappalainen, T., Gutierrez-Arcelus, M. & Dermitzakis, E.T. Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genet.* **7**, e1002144 (2011).
75. Degner, J.F. *et al.* DNaseI sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
76. Gaffney, D.J. *et al.* Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol.* **13**, R7 (2012).
77. Wright, F.A., Shabalin, A.A. & Rusyn, I. Computational tools for discovery and interpretation of expression quantitative trait loci. *Pharmacogenomics* **13**, 343–352 (2012).
78. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
79. Westra, H.-J. *et al.* Systematic identification of *trans* eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
80. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14–24 (2014).
81. Fairfax, B.P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).
82. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

© 2015 Nature America, Inc. All rights reserved.